DOCUMENT RESUME

ED 298 128 SP 030 567

AUTHOR Wiersma, William
TITLE The Alabama Career Incentive Program: A Statewide Effort in Teacher Evaluation.
PUB DATE Jan 88
NOTE 18p.; For related documents, see SP 030 565-566.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Observation Techniques; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; *Measurement Techniques; Teacher Behavior; Teacher Effectiveness; *Teacher Evaluation; Testing Programs; *Test Validity

ABSTRACT

This paper summarizes the Alabama Career Incentive Program, a statewide effort to evaluate more than 30,000 teachers including specialists in guidance, speech therapy, and so on. The evaluation was based on 10 classroom competencies and five outside-the-class professional competencies. A low-inference instrument measured competencies by using different sets of behaviors from the 117 item instrument. The professional competencies were evaluated using a high-inference instrument--the Evaluator Questionnaire. A statewide norming study involving all teachers and principals took place in the spring semester, 1987. The Classroom Observation Record was shown to measure the competencies reliably; the Evaluator Questionnaire data was not usable in discriminating among teachers due to the exceptionally high scores which nearly all teachers received. Elementary and secondary teachers differed significantly on several of the competency measures; no significant difference was found between special education teachers and their counterparts at the same level. A list of the competencies and a general account of the scoring procedures is given. (Author/JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298128

THE ALABAMA CAREER INCENTIVE PROGRAM:

A STATEWIDE EFFORT IN TEACHER EVALUATION

by

William Wiersma

January, 1988

Abstract

This paper summarizes the Alabama Career Incentive Program.

The Alabama Career Incentive Program was a statewide effort to evaluate more than 30,000 teachers including specialists in guidance, speech therapy, and so on. The evaluation was based on ten classroom competencies and five outside-the-class professional competencies. The instrument used for classroom observation was a low-inference instrument as described by Medley, Soar, and Coker. Competencies were measured by using different sets of behaviors from the 117 item instrument. A complex computer program was developed for generating scores and a teacher performance profile. The professional competencies were evaluated using a high-inference instrument - the Evaluator Questionnaire.

A statewide norming study involving all teachers and principals took place in the spring semester, 1987. The Classroom Observation Record was shown to measure the competencies reliably; the Evaluator Questionnaire data was not usable in discriminating among teachers due to the exceptionally high scores which nearly all teachers received.

Elementary and secondary teachers differed significantly on several of the competency measures; no significant difference was found between special education teachers and their counterparts at the same level. A list of the competencies and a general account of the scoring procedures is given.

SP030567

Original available from:

Don Taebel,
Assistant Project Direct
Teacher Evaluation Project
Georgia State University
Atlanta, GA 30303

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Coker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

THE ALABAMA CAREER INCENTIVE PROGRAM: A STATEWIDE EFFORT IN TEACHER EVALUATION

Introduction

Within the past decade or so there have been numerous and persistent attempts at large-scale assessment and evaluation of teachers. Undoubtedly, some of these attempts have been motivated by reports such as that of the National Commission on Excellence in Education (1983). But much of the motivation came from within the profession possibly from such factors as the competency-based education movement and the pressure to recognize excellence in teaching for raising the salaries of teachers. Whatever the reasons, the evaluation of teaching is a complex task, and when implemented at a state level, many issues and procedures, some difficult to deal with, are involved.

One approach to evaluating teachers' knowledge, not necessarily how they perform in the classroom, is through paper-and-pencil tests. The Texas Examination of Current Administrators and Teachers (Shepard and Kreitzer, 1987) is an example of such a test. This test of basic literacy was intended to identify incompetent teachers and to provide political leverage for raising salaries. But few people would argue that testing alone can provide an adequate assessment of teacher performance. Measures of actual performance of teaching are necessary. That approach has been used on occasion, an example being the Florida Performance Measurement System (Peterson, et al, 1986). This report provides an overview of another state-wide effort in the evaluation of teachers through observation; the Alabama Career Incentive Program (CIP).

Scope of the CIP

The enabling legislation for the CIP was passed in the State of Alabama in May, 1985. The law provides for an incentive-based, merit pay plan for teachers; based on their classroom performance and other professional activities.

The assessment of teachers is based on both professional judgment of best practice and the research on teacher effectiveness. The legislation established a working committee of teachers, school administrators and community leaders to design the evaluation program for the state's 42,500 instructional personnel. In terms of numbers of teachers impacted, the intent of the legislation was to reach all the state's teachers and supportive personnel such as guidance counselors.

The working committee formulated fifteen competencies representing the most important teaching activities, based on the Committee's interpretation of research results and practice. The competencies were then defined in terms of behaviors and activities observable in classroom performance and overall professional work. Although the competencies are limited in number, it was believed that they provided comprehensive coverage of teaching performance while keeping the competencies to a managable number. The specific competencies are:

1. Presenting organized instruction
2. Using teaching/learning materials
3. Providing for practice and application
4. Monitoring student achievement
5. Using monitoring data
6. Managing classroom time
7. Maintaining student behavior
8. Knowledge of subject matter
9. Maintaining a positive atmosphere
10. Communicating clearly and effectively
11. Preparing for instruction
12. Evaluating student achievement
13. Assuming responsibility for development or self and profession
14. Performing professional responsibilities
15. Human relations.

For each competency several indicators of the competency were developed with corresponding behaviors to be used in the assessment. An example of an indicator for Competency 7 is, "uses disciplinary strategies." Because of space limitations the indicators will not be listed here.

Measuring Performance

There are various strategies and approaches to the observation of teaching and Evertson and Green (1986) have provided a comprehensive summary of observation as a form of inquiry. The approach used for the CIP was that of a "sign system," in which samples of behavior or processes are taken in a specified time period.. Sign systems of observation have been classified as low-inference and high-inference depending on the extent of inference required of the observer in coding the teaching performance. Low-inference observation typically requires checking the occurrence of behaviors; high-inference observation some type of rating of performance, for example, on a scale of 1 to 5.

There are some dilemmas in the measurement of teacher performance caused by the differences in low-inference and high-inference observation. Dickson and Wiersma (1984) found that when the same competencies were measured using both high-inference and low-inference observation inventories, the correlations between the scores were disappointingly low, most around zero magnitude. This result indicates that measurement is highly instrument dependent. A disadvantage of high-inference observation is that measurement and evaluation are the same process. Low-inference observation is limited in discrimination power when measuring the occurrence of teacher behaviors that take place outside the classroom; essentially all behaviors occur in some form. For the CIP it was decided to develop a measurement-based inventory for observation. The approach is described in Medley, Coker and Soar (1984).

Four instruments were developed for measuring teacher performance; (1) the Classroom Observation Record (COR), (2) the Evaluator Questionnaire, (3)

the Pre-Observation Data form, and (4) the Post-Observation Data form. The latter two forms deal with information about class characteristics, lesson objectives, etc. The COR and the Evaluator Questionnaire form were designed to measure the 15 competencies of the Alabama teachers.

The COR is a low-inference inventory which provides a record of classroom interactions. It provides scores on the first 10 competencies listed earlier. There are two sections of the COR -- Section A consists of 69 items which describe specific interactions between teacher and students relating to instructing, questioning, responding, and managing. Section B which has 48 items focuses on the general teaching behaviors and the classroom context. The COR measures teaching performance taking place in the classroom and it does not measure supportive activities such as planning taking place outside the classroom.

The supportive activities, reflected in the final five competencies listed, were measured by the Evaluator Questionnaire. This form was used by the principal to measure professional activity such as professional service in the school system and the community, and it is completed in a conference between the principal and teacher.

Two classroom observations using the COR of at least 45 minutes each were conducted for each of over 30,000 teachers in Spring 1987. An observation lasted 45 minutes or an entire class period whichever was longer. This way a minimum of nine COR records were obtained since the observation for a single record is conducted in five minutes. Classes were observed in the teacher's major area of certification or in the area in which the teacher taught for the major portion of the day. The first visit was scheduled in advance by the teacher and observer. The second visit was scheduled by the observer, announcing the date but not the time of day for the observation. Near the end of the semester the principal and teacher completed the Evaluator Questionnaire covering the relevant activities of the entire semester.

As indicated in the description above, the CIP was an educational measurement effort of extensive scope, not only because of the large number of teachers involved but also due to the number of instruments and the data management procedures. CIP coordinators, principals, observers and teachers were in the field and all the data were sent to the project staff at Georgia State University. Data management procedures generally went smoothly with the observation forms being scored by optical scanners and placed directly onto a tape.

Results

The CIP produced a massive data bank, one that has the potential for countless analyses. The initial intent of the CIP program was to improve student learning and to provide a basis for a merit pay plan. But the data bank also provides a rich resource for research on teaching. Not only did the observation involve a large number of teachers, but it covered the spectrum of teaching, grades K-12.

Scoring Observation Data

Before considering specific results it is informative to comment on the scoring procedures. The raw data of the COR, as is the case with low-inference observation, consist of "bubbles" filled-in on the observation form. These marks need to be converted to scores with quantitative meaning, at least approaching interval scale measurement. Scoring involves two steps; (1) developing a scoring key and (2) normalizing the scores.

The items of the COR represent teacher behaviors and a cluster of the items is aggregated to provide a score for a specific competency. Each competency has its own cluster of items and the items in a cluster represent behaviors indicative of the competency. A cluster of items is called a scoring key, thus there are as many scoring keys as there are competencies. The number of items in the scoring keys vary across competencies.

The normalizing process of scoring involves two steps. The first step is to obtain the mean frequency per visit on each item and the item scores are normalized so that they can be added to generate a composite score for the competency. Then the second step is to compute an average score for the competency across all teachers and transform this score to a mean of 50. The distribution of all scores on the competency is transformed into a distribution with a standard deviation of 10. Thus, for each competency is obtained a distribution of scores, normally distributed, with a mean of 50 and a standard deviation of 10. Any teacher's score on a competency is then interpreted relative to the mean of 50. Clearly, this is a norm-referenced interpretation relative to whatever group of teachers has been used in the normalizing process.

The Evaluator Questionnaire has indicators for each competency. Indicators describe teacher activities indicative of the competency. Scoring of the Evaluator Questionnaire is straightforward, the codes for the indicators which simply indicate the presence or absence of the activities, are summed.

The validity and reliability of measurement is always a concern with programs such as the CIP. Essentially, at this point, validity is content validity. The development of the scoring keys for the COR, which was an extensive process involving successive approximations and refinements, is based on the content of the items as it is believed to relate to the competencies being measured. But reliability and the power to discriminate are more technical matters than validity.

Reliability of the COR refers to the reliability of individual competency scores. Thus, for a given group there are as many reliability coefficients as there are competencies. Cronbach's coefficient alpha, which is an internal consistency coefficient, was used to estimate reliability using 10% samples of the population data. Reliability coefficients for the 10 competencies measured by the COR varied considerably across competencies although they were highly consistent across groups of teachers identified by grade level taught. Most

reliability estimates were in the .65 to .90 range. The reliability of the competency scores from the COR was considered to be adequate.

The five competency scores obtained from the Evaluator Questionnaire had high reliability coefficients, four of the estimates being .95 or greater. This result was due to the indicators having extremely high percentages of "yes" responses indicating that the activities were done by the teacher. Because of the pattern of responses, the Evaluator Questionnaire had practically no discrimination power, that is it did not discriminate among teachers. Competency scores generated through the normalizing process would be meaningless and in a few cases misleading. Most teachers had scores at the mean of 50, and the very few teachers who received a "no" on only one or a few indicators had dramatically low scores, scores in the 20s or 30s. The conclusion about the Evaluation Questionnaire was that it needs revision either in its content, its data collection procedures or both. The remainder of this description is limited to results obtained by using the COR.

Profiles Based on the Ten Competencies Measured by the COR

As described in the scoring procedures, a score for each teacher was obtained on each competency measured by the COR. For any specified group of teachers used as a norm group, half of the scores on a competency will be above the mean and half below the mean. Scores for a given teacher may vary considerably across the competencies. If the 10 competency scores were combined into some type of composite score much of the information about the teacher's performance would be lost. A profile of a teacher's 10 scores seems to be the most appropriate and informative manner to describe the performance. Such a profile is similar to a profile of subtest scores in elementary school subjects and skill areas.

Profiles can be developed for individual teachers, in which case they would be useful for diagnostic purposes, or for groups of teachers, useful for comparison

purposes. Since the distributions are normalized, individual scores readily can be converted to percentiles.

Competency Scores for Six Sub-Populations

Profiles can take various forms; they can consist of tabled values such as individual scores or means of groups, or they can be displayed graphically. The latter shows differences more dramatically but if too many groups are included the profiles become confused. The population of Alabama teachers was divided into six subpopulations; elementary, middle school and high school teachers and then each of these divided into regular and special education teachers. The means for these six subpopulations on the 10 competencies are given in Table 1. The mean of each competency for the entire population of

Table 1 about here

teachers was established at 50. So going down a column provides the profile of mean competency scores for that subpopulation. For example, the regular elementary teachers scored above the overall mean of 50 on seven of 10 competencies. The lowest score for this subpopulation was on competency 9, a score of 47.4. Even this score is only about one-quarter of a standard deviation below the mean.

Inspection of the results of Table 1 reveals that differences between levels (elementary, middle and high school) are greater than the differences between regular and special education teachers within levels. Figure 1 contains two sets of profiles. In the upper part, the profiles for regular teachers are

Figure 1 about here

given for the three levels. The lower portion contains the profiles of elementary regular and special education teachers. The differences are considerably greater in the upper portion.

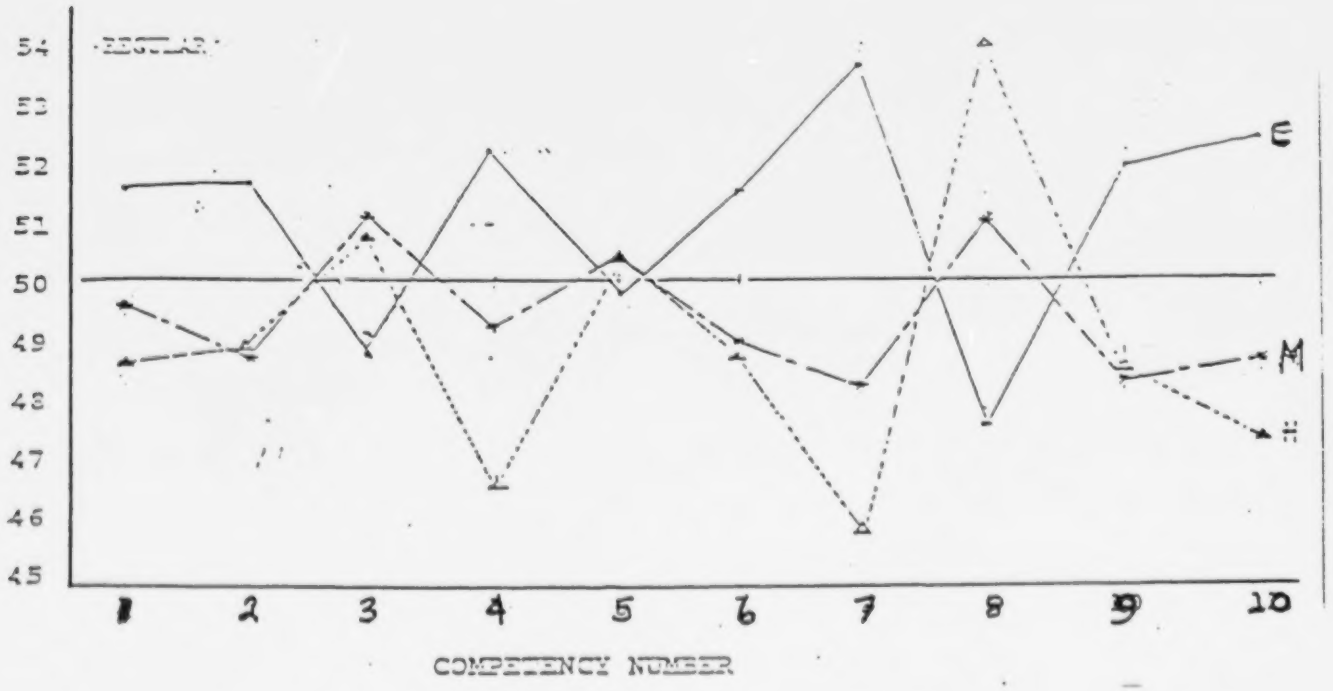
Table 1

Mean Competency Scores for Six Teacher Subpopulations

Competency Number	Teacher Subpopulation					
	Elementary		Middle		High School	
	Reg	SPE	Reg	SPE	Reg	SPE
1.	51.5	49.4	49.6	47.4	48.4	47.6
2.	51.6	52.8	48.7	48.4	48.5	47.6
3.	48.9	49.3	51.0	51.8	50.6	52.7
4.	52.2	52.1	49.4	50.8	46.5	49.6
5.	49.9	50.2	50.1	50.0	50.1	50.2
6.	51.5	50.6	49.0	48.0	48.8	47.6
7.	53.5	52.8	48.3	48.5	45.8	46.2
8.	47.4	47.1	51.0	49.4	53.8	51.6
9.	51.7	52.7	48.3	49.1	48.4	48.3
10.	52.3	51.5	48.7	48.8	47.4	47.3

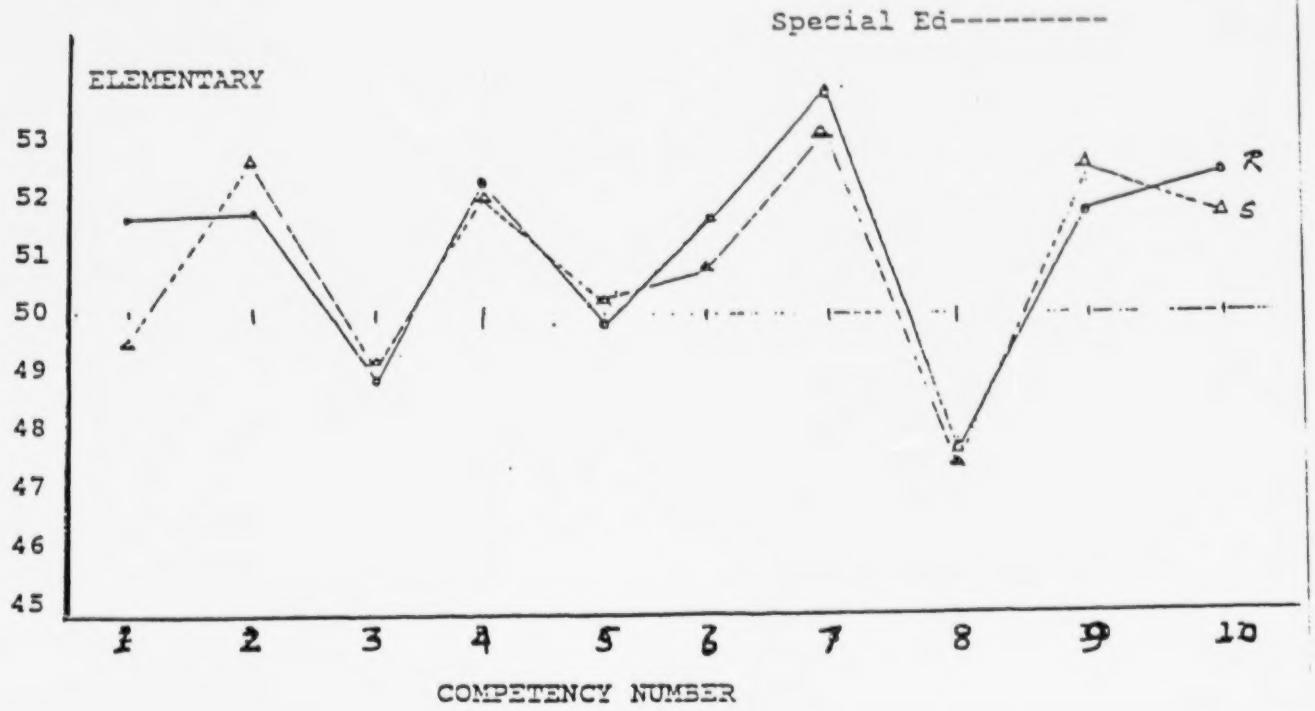
Figure 1

Profiles of Mean Competency Scores of Selected Teacher Subpopulations



Regular Ed_____

Special Ed----------



Consider competencies such as numbers 4, 7, 8 and 10, those competencies showing the greatest differences between the means of regular teachers. These competencies are:

4. Monitoring student achievement
7. Maintaining student behavior
8. Knowledge of subject matter
10. Communicating clearly and effectively

The elementary teachers had the high means on three of the four competencies. Only on competency 8, "Knowledge of subject matter," did the secondary teachers have a markedly higher mean than the other two groups. Does this mean that the elementary teachers are better teachers, so to speak? Possibly, but there are alternate explanations of the results. It may be that the difference in competency scores reflects differences in the nature of teaching at these levels. High school teachers may not have the need to monitor student achievement in as continuous a manner as elementary teachers. Consequently the high school teachers would not exhibit the monitoring behaviors necessary to attain high scores on this competency as would the elementary teachers. Because high school courses are more subject oriented it may be easier, even necessary, for high school teachers to demonstrate a knowledge of subject matter than teachers at lower levels.

The overall configuration of the upper portion of Figure 1 supports the idea of differences in competency scores being at least in part, due to differences in the nature of teaching at various levels. The means of the middle school teachers generally position between the means of the other two levels, showing a consistency of differences related to the levels taught.

If performance on the competencies is highly impacted by the nature of teaching at various levels rather than the skill of the teacher, does this make competency scores useless for differentiating between teachers? Not at all, but this result does indicate that a variety of norm groups are necessary. On the

basis of these results, it does not seem appropriate to use the entire population of Alabama teachers as the norm group for individual teachers. This causes no problem, because adequate data are available for establishing norms for any of the six subpopulations of teachers. Interpreting results within subpopulations would be more meaningful, and less susceptible to erroneous conclusions, than using the data from all teachers as a single population.

Correlations Among Competencies

There is, of course, a plethora of results and even entire studies contained in the CIP data base. Space limitations prohibit any detailed discussion, but one additional result merits mention. One of the difficulties with teacher evaluation based on observation has been the lack of distinction among teaching competencies as evidenced by high correlations among competency scores and what might be called a "global" impression of teaching. Measurement-based evaluation, using low-inferenced observation appears to overcome this difficulty to a large extent. The correlations between the scores for the 10 competencies measured by the COR were computed. Of the 45 correlation coefficients in the matrix, only two were .50 or higher and 26 were .20 or lower. This result indicates that teaching as measured by the COR is multi-dimensional.

A second analysis approach which involves correlation is factor analysis. Rather than factor analyze competency scores which are only 10 in number, it was decided to analyze scores on indicators. Indicators are subsets of behaviors (items of the COR) which are indicative of specific competencies. When factor analyzing indicator scores, 16 indicators had factor loadings of .50 or greater across four somewhat distinct factors. The factors had eigenvalues ranging from 3.75 to 1.78. Names of factors are, of course, arbitrary, but based on the indicators loading on the factors, appropriate names of the factors are:

Teacher and student interaction in instruction

Guided practices with teacher elaboration

Classroom management

Use of materials in instruction

The results of this factor analysis were consistent with findings of Dickson and Wiersma (1984), in which scores on 22 competencies were factor analyzed. The 22 competencies had been measured using a low-inference inventory, the Classroom Observations Keyed for Effectiveness Research (COKER). The COKER is an inventory similar to the COR, with about three times as many items. Dickson and Wiersma found six factors with eigenvalues ranging from 6.77 to 1.01. Overall, the relative independence of the competency scores and the consistent factor analysis results across independent studies provide evidence that low-inference observation measures teaching as multidimensional with identifiable constructs.

Conclusions

The Alabama CIP study has generated a large data base and this report has only touched upon the results and possible conclusions. Much information presently lies unrevealed in the data base. With a study of this type there is a tendency to focus on the quantitating results. These results are certainly important, but also of importance is the process of the project.

The CIP is an effort that involves thousands of teachers, principals, supervisors and project staff. The data collection and management are extensive, seemingly overwhelming processes. However, the program has been shown to be feasible, successfully involving teachers in structuring the program. Teachers and principals have become aware of measurement concepts and problems. They have been required to think of teaching relative to competencies that are demonstrated by specified behaviors observed in the classroom. All of this highlights the importance of the observation process. It can be concluded that "the process is at least to some extent, the product."

The COR was shown to be a manageable and valuable observation inventory. It measured 10 competencies demonstrated in the classroom with adequate reliability. The competency scores showed teaching to be a multidimensional skill or process. Observation data can be used in a variety of ways; probably one of the most useful being profile analysis for individuals or groups of teachers. The COR clearly has discrimination power for measuring teachers on the competencies.

The use of the Evaluator Questionnaire was disappointing, primarily because it has no discriminating power. For the purposes of the project, discriminating power is paramount. Further development of the Evaluator Questionnaire is necessary and it is likely that a major revision of this form will be required before it produces useful data.

Results of profile analysis of the competency score means for subpopulations of teachers by level taught were discussed as an illustration of information obtained using the COR. The subpopulations did separate on many of the competency scores, indicating different teaching performance. Differences may be due to variations in the nature of teaching at different levels and they may reflect differences in training programs. It is possible that certain subpopulations simply are more proficient on some competencies than others. The massive data base of the CIP has the potential of a definitive study on the nature of teaching.

There are several issues of the Alabama CIP that have not been raised in this report because of space limitations. As an example, one such issue is methods for setting standards of teaching proficiency. Defining criteria will be subjective to a large extent, but various methods can be simulated to obtain models of results. Whatever method is used, judgments will likely be made using subpopulation norm groups rather than the entire population of teachers as the norm group. An inspection of the data shows that as grade levels increase the

mean scores on some competencies decrease. But this result may be due to variations in the nature of teaching rather than less competent teachers at higher levels.

To date, the Alabama CIP has been a major effort in the measurement and evaluation of teaching. Much work remains to be done, even for the complete analysis of the present data base. The procedures and process for measurement have been defined and established. The process and resulting data of the COR have been demonstrated as feasible and valuable. The CIP is a major effort in the evaluation of teaching, intended to upgrade the profession and improve the teaching in the schools.

References

1. Dickson, G.E. and Wiersma, W. (1984). Research and evaluation in teacher education: Empirical measurement of teacher performance. Toledo, OH: The Center for Educational Research and Services; The University of Toledo.

2. Evertson, C.M. and Green, J.L. (1986). Observation as inquiry and method. In M.C. Wittrock (Ed.) Handbook of research on teaching, 3rd ed. New York: MacMillan, pp. 162-213.

3. Medley, D.M., Coker, H., and Soar, R.S. (1984). Measurement-based evaluation of teacher performance: An empirical approach. New York: Longman, Inc.

4. National Commission on Excellence in Education (1983). A nation at risk. Washington, D.C.: Government Printing Office.

5. Peterson, D., Micceri, T., Kromrey, J., and Smith, B.O. (1986). Teacher evaluation for advancement or retention: Issues and application. In W. Wiersma (Ed.) Proceedings of the April 1986 conference: Issues in certification, advancement and instruction. Toledo, OH: The Center for Educational Research and Services; The University of Toledo.

6. Shepard, L.A. and Kreitzer, A.E. (1987). The Texas teacher test. Educational Researcher, 16(6), 22-31.